Huan Zhao. Data Visualization and Analysis about Physicians' Payment Data. A Master's Paper for the M.S. in IS degree. Apr, 2018. 39 pages. Advisor: Arcot Rajasekar

In this paper, I analyze the relationship between the number of patients that have been identified with a specific type of disease and the physicians' total charge amount. I hold the assumption that the more states within a region that show statistical significance in the relationship, the more serious the disease in that region will be. But after doing a detailed analysis and visualization, I found the assumption to be invalid. However, I found the incidence rate for some of the diseases were related to geographical location. Therefore, I did a comprehensive research to find the reason for this relationship from different perspectives as well as provide some recommendations.

Headings:

Disease

Medicare payment data

Visualization

DATA VISUALIZATION AND ANALYSIS ABOUT PHYSICIANS' PAYMENT DATA

by
Huan Zhao

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2018

Approved by

_____

Arcot Rajasekar

Table of Contents

1. Introduction

Every year, CMS (the Centers for Medicare and Medicaid Services) releases datasets containing information on Medicare payments made to physicians and other providers. Though this is an important achievement in promoting greater health system transparency, much hidden information has not been explored.

There are some suggestions about the potential use of the dataset. For consumers, the release of these data could eventually facilitate comparison among individual physicians, of types of services delivered, and payments received (Blum, 2014). “If presented in user-friendly ways, and paired with information on quality of care, the data could help consumers choose physicians who deliver the highest-quality care” (Patel, Masi & Brandt, 2014). This means if we could pair the data with the quality of care, it is easy to find the physician that is most suitable to our need. Some researchers also suggest patients could use the dataset to find physicians with higher bills, then “construct provider networks and insurance products” to benefit both the payer and physicians as well as decrease the healthcare cost (Patel, Masi & Brandt, 2014). For this reason, finding the hidden information behind the dataset is critical for providers, physicians, insurance companies and other interested parties.

In this project, I explored this dataset in a new way. I visualized the relationship between the percent of beneficiaries identified with different types of disease and the physician's total submitted charge amount. By analyzing whether there is statistical significance in the relationship, I am trying to figure out whether the positive correlation means the prevalence of disease in that region.

2. Literature review

In order to explore what I could do with the dataset, I did a broad literature review, including the use of the payment dataset, the cost for different types of diseases and disease distribution according to geographical location.

The dataset from CMS is valid, they take "data integrity very seriously and took swift action after a physician reports a problem" (Ornstein, 2014). Hence, we could make sure the data is accurate and update, and would be a good resource to do some research and analysis. Patel, Masi and Brandt did a research about the potential use of the dataset. They found "the majority of research to date has been conducted by the media and has focused on specific high-cost providers and specialties, procedures, and drugs". From this we could know some organizations have paid attention to the dataset, but most of the research are still in the initial stage, since most of them focus on data with specific character, not the whole dataset. The data "could tell us whether the variation in utilization was greater among individual providers or between health care markets" (Patel, K. P., Masi, D. M., & Brandt, C. B. (2014)). Therefore, patients could compare their provider with other physicians to find if there is great difference in terms of drugs, fees and treatment. With the dataset, patients will be able to see the total Medicare payments a physician received and determine which physician to choose based on claims

from Medicare and the rank of the physician in an area (Ogara, 2014). The “disclosure of the individual physician payment has been advocated as a powerful tool to control healthcare cost and to improve the delivery of care” (Steinbrook, 2014). However, there are also some opinions that this is not good for physicians’ privacy (Steinbrook, 2014).

Physicians’ malpractice payments kept growing from 1991 to 2003, and this is consistent with increases in the cost of healthcare (Chandra, 2005). Different types of treatments may have different effects on the profitability and the amount of submitted charge (Dugel &Tong, 2011). There are many factors which will affect billing charges of medical providers, such as the cost of living, the “age of the provider, specialty, and the percentage of patients under a managed care programme” (Canavos & Brown, 2004). Most geographic cost variations in Medicare are due to health differences (Reschovsky, Hadley & Romano, 2013).

Some diseases are more expensive than others. The top ten most expensive diseases include HIV, cancer, transplant, stroke, hemophilia, heart attack, coronary artery disease, neonate, end-stage renal disease, and respiratory failure on ventilator (Whelan, 2012). The top 5 medical services expenditures in the US by disease category in 2013 were ill-defined conditions, circulatory system, musculoskeletal, respiratory and endocrine. What’s more, the cost of these diseases kept increasing from 2000 to 2013(“How much”, 2017). “America spends nearly $800 billion fighting the most common neurological disorders, which include Alzheimer's disease, Parkinson's disease, epilepsy, multiple sclerosis” (Lisa, 2018). With the growth of the older population, the cost of neurological

disorders is predicted to increase a lot (Lisa, 2018). National Cancer Institute estimates that by 2020, the cost for cancer will be more than $150 billion (Lisa, 2018). Mental disorders cost about $200 billion, with nearly fifty percent of Americans having mental disorder (Lisa, 2018). According to the American Diabetes Association, the cost for diagnosing diabetes was $174 billion in 2007, and increased to $245 billion in 2012. During these five years, the cost increased nearly fifty percent (Burns, 2018). More than sixty percent of the "cost is paid by the government insurances such as Medicaid and Medicare", which verified the effectiveness of our dataset from another perspective (Burns, 2018).

The cost for disease also depends on various factors. For example, Whelan mentions that the cost for cancer depends on treatment factors. If the patient doesn't take radiotherapy and chemotherapy treatment, it may cost no more than $20,000 a year. "This would apply to 40% of total cancer patients" (Whelan, 2012). However, if the patient needs a surgery, it costs much more. Age is another factor that may affect the cost. Older women usually spend more than men and young people. For example, in 2013 "women ages 85 and older spent the most, and about 58% "occurred in nursing facilities, while 40% was expended on cardiovascular disease, Alzheimer's disease" ("Diabetes ", 2016).

3. Analysis for Cost and Incidence

3.1 Dataset Introduction

The dataset I used is "Medicare Physician and Other Supplier National Provider Identifier (NPI) Aggregate Report" for the calendar year of 2015. It contains comprehensive information about providers, patients and payments. Including geographic information, such as location, demographic information, such as gender, race, and health characteristics, such as proportion of patients in a specific age group.

I downloaded the dataset from CMS official website, the original dataset is an excel file. It has more than one million rows and seventy columns. Since my research goal is to find the relationship between provider's total submitted charge amount and the proportion patients identified with a specific type of disease. I kept the data about the percentage patients have identified with 16 types of disease and total submitted charge amount as well as physicians' geographical location.

There are 31 states in our dataset, including Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York, Pennsylvania, Illinois, Indiana, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota, Delaware, Florida, Georgia, Maryland, North Carolina,

South Carolina, Virginia, District of Columbia, West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma, Texas, Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, Alaska, California, Hawaii, Oregon and Washington. The 16 types of diseases including Atrial Fibrillation, Asthma, Cancer, Heart Failure, Chronic Kidney Disease, Chronic Obstructive Pulmonary Disease, Depression, Diabetes, Hyperlipidemia, Hypertension and Ischemic Heart Disease, Osteoporosis.

3.2 Data visualization

I divided the original dataset into 31 small datasets according to physicians' state. Then I cleaned each of the small dataset by deleting rows which contain invalid data. After cleaning, the data size for each state ranges from two thousand rows to more than 25 thousand rows.

For data visualization part, I referenced what we have learned in the course of Information Visualization, using Tableau to draw the graph. For each dataset, I made sixteen scatter plots for sixteen types of diseases. X axis represents for the proportion that patients have identified with that disease, y axis represents for the physician's total charge, each point on the graph represents one line of data. Then Tableau will calculate the p-value for each type of disease. The threshold value is set to 0.05, that is, if the calculated p-value is less than 0.05, we consider there is statistical significance between the disease and total charge amount, if the value is more than 0.05, we consider this is

due to random error. According to the p-value, we divided the result into two parts, statistically significant and not significant.

Below are some examples of the visualization result for Oklahoma. From it we could know in this state, the total charge amount is related to the number of patients with Atrial Fibrillation, Alzheimer's disease or Dementia, Asthma and Cancer. However, Hyperlipidemia and Rheumatoid Arthritis / Osteoarthritis don't show statistical significance with total submitted charge amount.

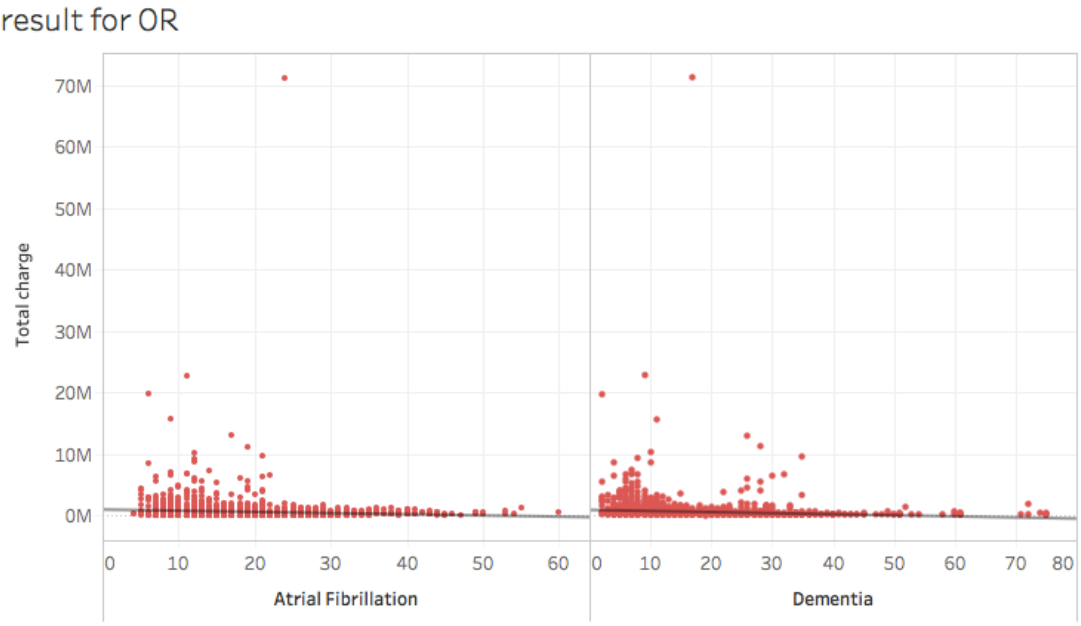

Figure 1: Example of Oklahoma, diseases show statistical significance

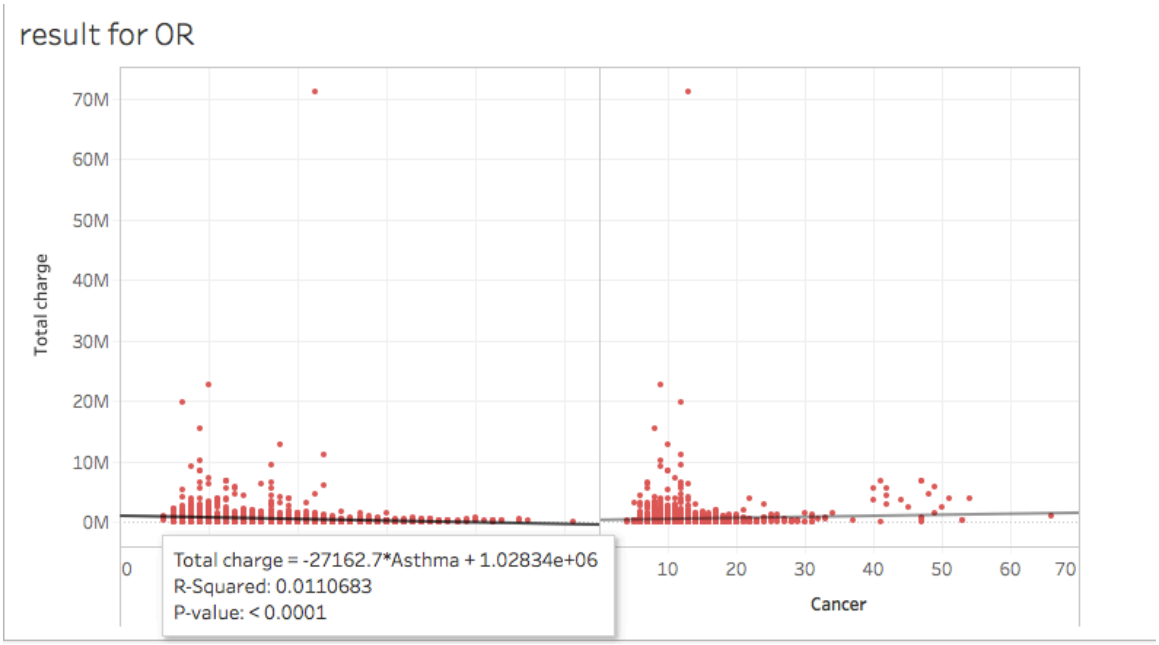

Figure 2: Example of Oklahoma, diseases show statistical significance

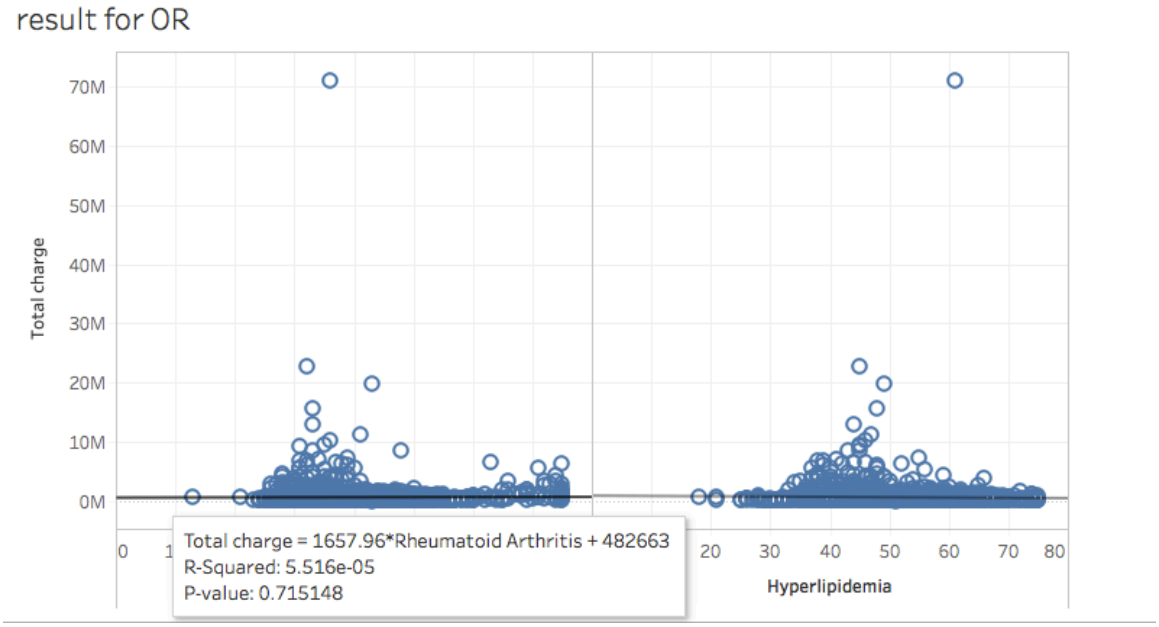

Figure 3: Example of Oklahoma, diseases show statistical insignificance

A linear trend model is computed for Total charge given Ischemic Heart. The model may be significant at p <= 0.05.

Model formula:	(Ischemic Heart + intercept)
Number of modeled observations:	2417
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	2415
SSE (sum squared error):	7.91376e+15
MSE (mean squared error):	3.27692e+12
R-Squared:	0.0028549
Standard error:	1.81023e+06
p-value (significance):	0.0086047

A linear trend model is computed for Total charge given Obstructive Pulmonary. The model may be significant at p <= 0.05.

Model formula:	(Obstructive Pulmonary + intercept)
Number of modeled observations:	2417
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	2415
SSE (sum squared error):	7.88551e+15
MSE (mean squared error):	3.26522e+12
R-Squared:	0.0064142
Standard error:	1.80699e+06
p-value (significance):	< 0.0001

Figure 4: a model for each graph

Individual trend lines:

Panes		Line		Coefficients				
Row	Column	p-value	DF	Term	Value	StdErr	t-value	p-value
Total charge	Hyperlipidemia	0.207396	2415	Hyperlipidemia	-5331.73	4227.85	-1.2611	0.207396
				intercept	815925	215821	3.78056	0.0001603
Total charge	Dementia	< 0.0001	2415	Dementia	-17598	3847.41	-4.57398	< 0.0001
				intercept	838415	73390.6	11.424	< 0.0001
Total charge	Asthma	< 0.0001	2415	Asthma	-27162.7	5224.65	-5.19895	< 0.0001
				intercept	1.02834e+06	99447.1	10.3406	< 0.0001
Total charge	Atrial Fibrillation	< 0.0001	2415	Atrial Fibrillation	-19305.8	4764.76	-4.05179	< 0.0001
				intercept	900306	94453.7	9.53172	< 0.0001
Total charge	Cancer	0.0151012	2415	Cancer	16896.8	6948.6	2.43168	0.0151012
				intercept	331765	96155	3.45032	0.0005695
Total charge	Depression	< 0.0001	2415	Depression	-23598.4	4090.62	-5.76889	< 0.0001
				intercept	1.27866e+06	131884	9.6953	< 0.0001
Total charge	Diabetes	0.0014627	2415	Diabetes	-13849.7	4347.52	-3.18565	0.0014627
				intercept	1.021e+06	153046	6.67121	< 0.0001
Total charge	Heart Failure	< 0.0001	2415	Heart Failure	-12842.6	2772.1	-4.63281	< 0.0001
				intercept	931698	90642.1	10.2789	< 0.0001
Total charge	Hypertension	0.0002298	2415	Hypertension	-16077.6	4357.88	-3.68933	0.0002298
				intercept	1.62976e+06	295576	5.51384	< 0.0001
Total charge	Ischemic Heart	0.0086047	2415	Ischemic Heart	-8223.58	3127.4	-2.62953	0.0086047
				intercept	859926	124297	6.9183	< 0.0001
Total charge	Obstructive Pulmonary	< 0.0001	2415	Obstructive Pulmonary	-14850.2	3761.03	-3.94845	< 0.0001
				intercept	879349	91672.2	9.59232	< 0.0001

Figure 5: general description for the dataset

I chose Tableau to visualize the data, because it is efficient and convenient. After connecting to data source, and selecting suitable model, I only need to drag the variable that needs to be visualized. Previously, I planned to visualize it with JavaScript and D3, then I found it is complex. I need to transfer the 31 files from CSV format to JSON

format. Since we have 16 variables, the size of each dataset will be 16 times larger after transformation. This will make the program runs very slow. In addition, the graph could only show the p-value, doesn't include the degree of freedom, standard error, t-value, and the function for each model. In terms of these reasons, I decided to use Tableau to visualize the dataset.

3.3 Result

After visualizing scatter plots for all the 31 states, I summarized a table for the overall result.

	A	B	C	D	E	F
1	Disease	West	South	midwest	Northeast	
2	Atrial Fibrillation	4/4	10/12	8/10	5/5	
3	Alzheimer's Disease or Dementia	4/4	11/12	10/10	5/5	overall
4	Asthma	4/4	11/12	10/10	5/5	overall
5	Heart Failure	4/4	10/12	10/10	5/5	
6	Kidney Disease	4/4	11/12	9/10	5/5	
7	Chronic Obstructive Pulmonary Disease	4/4	9/12	10/10	5/5	
8	Depression	4/4	11/12	10/10	5/5	overall
9	Schizophrenia/ other psychotic disorders	4/4	11/12	10/10	5/5	overall
10	Stroke	4/4	11/12	9/10	5/5	
11	Cancer	3/4	8/12	10/10	1/5	
12	Diabetes	4/4	8/12	8/10	2/5	northeast
13	Hyperlipidemia	1/4	0/12	4/10	1/5	south
14	Hypertension	4/4	4/12	5/10	2/5	west
15	Ischemic Heart Disease	4/4	8/12	5/10	4/5	
16	Osteoporosis	2/4	5/12	9/10	2/5	midwest
17	Rheumatoid Arthritis/ Osteoarthritis	0/4	3/12	2/10	0/5	(probability)
18						

Figure 6: Summary for the whole dataset

I divided the 31 states into four regions according to US Census Bureau to see if there is a general pattern for a specific type of disease in a region, including west, south, Midwest and northeast. In this table, data in the first column represents for the type of disease, and number in following columns are the result for each region. 4/4 means that there are four

out of four states in this region show statistical significance between number of patients identified with a specific disease and the physician's total charge. For the column F, overall means all four regions show statistically significance, but for Northeast/South/West/Midwest, it means that this region shows difference among the four regions. For example, for the disease of Diabetes, I found in northeast, the proportion of states which shows significance is less than fifty percent, whereas the proportion for other three regions are all much more than fifty percent. After researching the region which show difference from other regions in some disease. I didn't find something useful to explain the table. Even some regions show difference in some diseases, this is not related to the actual situation of the disease.

From the above results, it seems like the significance between the proportion of patients identified with a specific type of disease and physician's total charge amount doesn't indicate something meaningful. Like my previous assumption is that the significance means that disease is severe in that region. After talking with my advisor, I decide to research the prevalence for all the diseases contained in the dataset in 31 states, then choose some typical diseases which show significant geography prevalence to do further analysis.

4. Regional Analysis for Some Specific Diseases

I chose two diseases which show obvious geographical prevalence to do further analysis, including diabetes and cancer.

4.1 Diabetes

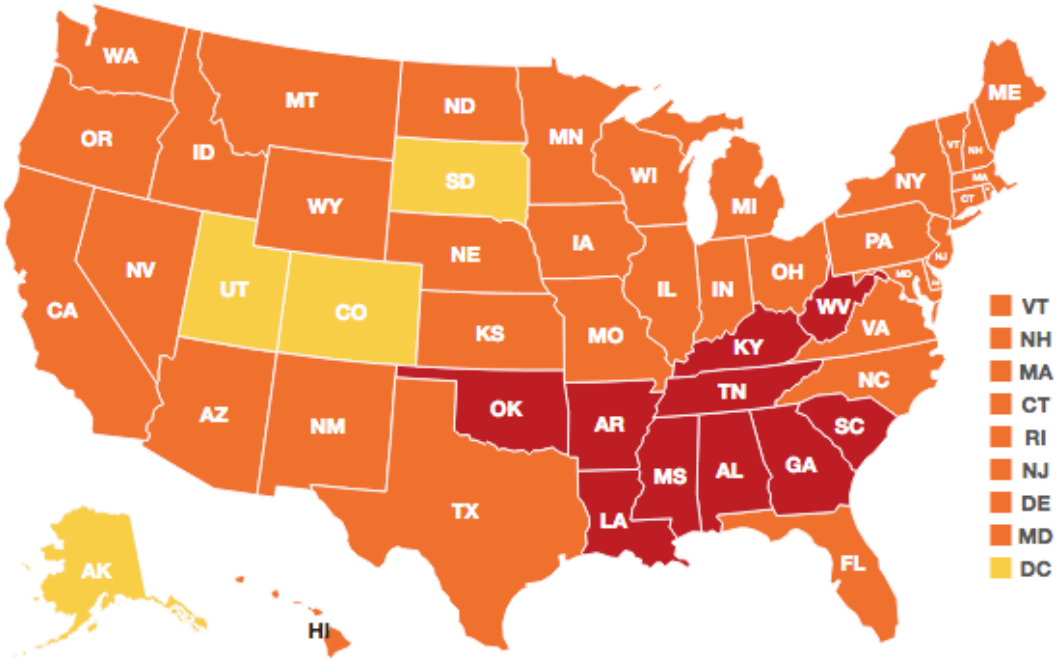

Figure 7: Diabetes rate for each state, from the state of obesity

This is a map from the State of Obesity, it shows the diabetes rate for each state in the US in 2016. With the increase of the incidence rate, the color for the state becomes darker. From the map we could know in 2016, all the top ten highest rates of type 2 diabetes states are in the south region, including West Virginia, Alabama, Mississippi, Arkansas,

Kentucky, South Carolina, Tennessee, Louisiana, Georgia and Oklahoma. And from a line graph for incidence rate about each state from 1990 to 2016 ("Diabetes in"), we could know that the diabetes rate generally kept increasing for all the states from 1990 to 2016. CDC scientists also identify a diabetes belt, which includes "664 counties in parts of Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, North Carolina, South Carolina, Ohio, Pennsylvania, Tennessee, Texas, Virginia and West Virginia" ("where in", 2015). It is obvious that almost all the states are in the southern region. In the nation, the rate of type 2 diabetes is about 8.5 percent, but in these areas, about 12 percent of people have type 2 diabetes ("where in", 2015).

After collecting information from some official websites, articles and researches, I found the reason could be summarized into different aspects, including diet structure, lifestyle, socioeconomic status and climate.

Lifestyle: About 30 percent of people in these areas have sedentary lifestyle, however the rest of the country is only 25 percent ("where in", 2015). It is well known that sedentary lifestyle is harmful to our health, and will increase the risk of many diseases, include obesity. Sitting for a long time without exercising is likely to result in the accumulation of fat. And when the intake calorie is more than the heat we consumed, it is easy to cause the obesity. Obesity is one of main courses of this diabetes. So that this unhealthy lifestyle is easily to result the increasing rate of diabetes. From recent data, 9 of top 10 states with the highest rate of obesity are in the southern region in 2016, and the average rate is about 33.11 ("Adult Obesity"). This should arise our attention. Another unhealthy

lifestyle is the poor-quality sleep. From a report about sleeping condition of the American (Charles, 2012), they found five southern states have the most people with sleep disturbances, and this probably result in the higher rate of obesity ("Hargens, 2013"), which increase the risk of diabetes indirectly.

Socioeconomic status: "Overall, the South has had lower percentages of high school graduates, lower housing values, lower household incomes". ("Southern", 2018). About 26 percent of people in the southern region have a bachelor's degree which is obvious lower than the rest of the country which is over 30 percent ("Southern", 2018). "Life expectancy is lower and death rates are higher in the South than in other regions of the United States for all racial groups" ("Southern", 2018). When the whole education level is low, this means people may not realize the importance of preventing the disease. Since they have less access to learn about the disease, for example, how it is formed, what it will bring to their further life, and how to prevent it. Especially diabetes is a chronic disease, it will not have significant influence on patients' life at the initial stage, therefore they probably will not pay much attention to it. Income is another important factor for the high incidence rate of diabetes. The income level decides which type of treatment patients could have. Some patients may have a chance to be cured, but consider for the cost, they may choose a less effective but cheaper way. People with lower income usually do not pay much attention to their lifestyle and diet structure, and they may not have regular examination, this potentially increases the probability to getting diabetes.

Environment: “extreme heat can affect our blood sugar control” (Hamaty, 2011). There are some studies find that high temperature will increase the probability of getting diabetes, and the influence of temperature could be divided into two aspects. Some researchers (Williams, 2017) suggested that the link could be attributed to a tissue in our body- “brown fat or brown adipose tissue” (Williams, 2017). The function of it is to keep the warm of our body. When the temperature is high, there “is less work that brown fat has to do to keep the body warm” (Williams, 2017). When the number of brown fat is less, the rate of getting diabetes is getting higher. Researchers also found that the increase of temperature will result in the increase of obesity (Williams, 2017), therefore the rate of diabetes is likely to increase. When the temperature is high, people tend to reduce the activity and take more cold drinks. This will increase calories intake and reduce the consumption, so that will increase the chance of obesity. Another theory is that the heat affects people with diabetes depends on “whether they’re well-hydrated and their activity level” (Hamaty, 2015). If the temperature is high, and patients do much activities which makes them profuse, they will lose much water. And this will lead to the rise of glucose levels. And if the blood glucose levels rise, it will lead to frequent urination, which makes the patient lose more water, the blood sugar level will probably higher, this is a negative feedback (Hamity, 2015).

Diet structure: When it comes to diabetes, eating is a topic that we couldn’t avoid. I found an interesting article about the diet structure of southern people, it mentions that they tend to have the unhealthy tradition of processing food. That is, some food is healthy originally, but in order to make it more flavor, people always add some additions to it,

which is possibly make it unhealthy (Toby & Alan). For example, adding chocolate chip to yogurt, adding jam to fruit, frying food in lard, this obviously will add much more calories and fat to the food. Biscuit is a representative of this. It is a popular cookie in the south, because of the good texture and flavor, some people like it very much, even eat it every day for the breakfast. But they may never consider the saturated fat bring by the lard and butter, which could be a course for diabetes.

4.2 Cancer

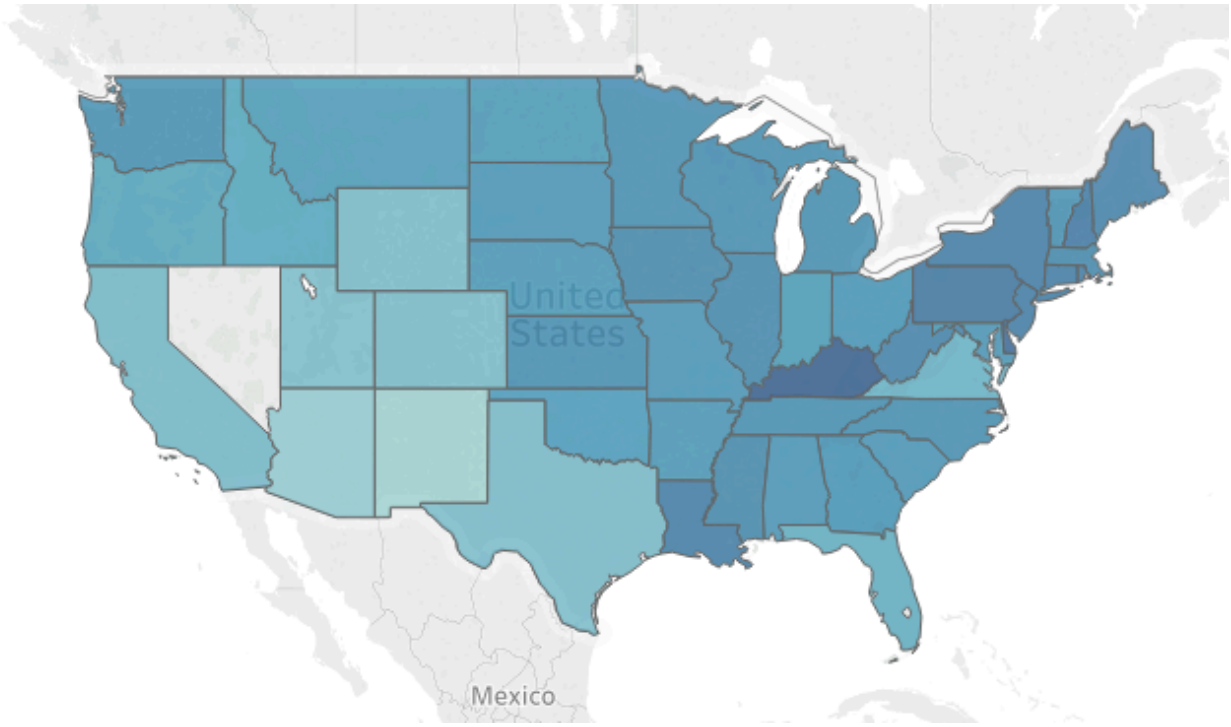

Figure 8: Age-adjusted incidence rates by cancer site (2010-2014)

After searching for the cancer incidence rate data from the website, I visualized this map by tableau. It represents the overall cancer rates in each state in US, range from 350.3 per 100,000 persons to 510.8 per 100,000 persons. From the graph we could observe that

geographical difference exists in the country, and the darker states concentrate in the northeast region.

Some articles and official website also validates this pattern. For example, a study from the U.S. Agency for Healthcare Research and Quality shows that "people living in the northeast are more likely to have brain cancer than those living in the South or the West. People in the Northeastern U.S. are one-third more likely than those in the South or West to be hospitalized for treatment of brain cancer or to have brain cancer when they are hospitalized for another illness or complication" ("Northeast", 2009). Another article points out the number of children diagnosed with cancer is much more than other areas ("Kid's", 2008). Some experts said this is because northeast part is developed, therefore more children have access to medical facilities, result in more diagnoses ("Kid's", 2008).

From CDC's data, eight out of eleven states in the northeast region are in the list of top 15 highest rates of cancer, including Massachusetts, Maine, Connecticut, New Hampshire, Rhode Island, Pennsylvania, New Jersey, New York. Among them, New York is one of the states with highest numbers of overall cancer cases. Common types of diagnoses in these states including female breast cancer, prostate cancer, lung cancer and bronchus cancers (Becker, 2018). The American Cancer Society predicts there will be more new cases in these states in following years (Becker, 2018). The mortality of breast cancer for women is also higher in the northeast than other areas in the US. What's more, another earlier study also found that "among white women aged 50 and older from the northeast, breast cancer mortality was 30 percent higher than for similar women in the

South and about 12 percent higher than for women in the Midwest and West" (Howe, 1995). Below is a table about the incidence rate for overall and specific cancer types. The data is from Centers for Disease Control and Prevention, link is provided in the reference. It is obvious that the rate in the northeast is higher than other regions.

	US	Western	Midwest	Southern	Northeast
Overall	436.6	401.1	430.1	432.3	471.8
Female breast cancer	123.9	120.3	126.3	119.8	134.2
Prostate cancer	95.5	85.5	96.7	96.4	104.7

Table 1: Cancer rate for overall US and each region

Accounting for the courses of cancer in the northeast part is difficult, since there are so many variables that may play a role ("Which U.S.", 2015). For example, age, race, socioeconomic status, lifestyle, pressure, gene inheritance, sun exposure and many other factors we may not familiar with ("Which U.S.", 2015). But after researching, I found some main courses which may explain why the incidence rate is higher than other regions.

More access to medical facilities: two articles have mentioned that more access to medical facilities potentially is a reason for the higher cancer rate ("Which" & "Kids'"). I think this makes sense, more access to medical facilities means more patients could be diagnosed earlier. Many big cities have excellent health care in the northeast region. For example, "More than a dozen hospitals in the Chicago area are nationally ranked for the quality of their medical care and superior staff" ("Advantages of"). In big cities like this, it is easy to find hospitals, specific clinics which fit your needs. And patients usually have

more choices, they can compare different hospitals and staffs then choose the best one. Even some of physicians may not be available, they can still choose others. Therefore, it's easier to get the resource and care we need in big cities. But living in rural areas means it always take more time and effort to visit a doctor, and sometimes the quality is not satisfied ("Advantages of"). Therefore, lower rate in other regions doesn't mean there must be less patients, some people may suffer from cancer but they don't know, especially for the type that doesn't have obvious reaction in the initial stage, such as the thyroid carcinoma.

Lifestyle could be the primary reason for the high cancer rates in the northeast. There are many big cities in the northeast such as New York, Boston and Chicago. People living in big cities tend to have a different lifestyle from other regions. For example, they usually have much more pressure from work and life, which will lead to smoking, drinking, sleeping disturbance and mental illness. And they are always in a hurry, they don't have much time for a good meal or sufficient sleep. Working to midnight is an ordinary thing for those work in big companies, flying around the world without adjustment of sleep is common. Even though, they face more challenge and competition every day. Having this lifestyle for a long time is likely to increase the risk of many different type of diseases such as asthma, heart disease and stomach disease. And these diseases will increase the risk of cancer instead.

Other research also mentions lifestyle is the main reason for the female breast cancer (Howe, 1995). Delaying marriage is very common in big cities. Women choose to marry

later because various reasons, such as education and personal motivation. For an ordinary woman, after getting the bachelor's degree, and work for some time, their age is around thirty. If they want to study further, it is possibly to postpone the time to have a child. More importantly, more and more couples decide not to have children. Both of these two have been validated as the cancer risk factors (Howe, 1995). Blot and McLaughlin had conducted a study, they found after twenty years old, the risk of having breast cancer increases seven to eight percent each year with the time of having a child (Howe, 1995). I made a table about the median marriage age in the US, we could find all the oldest-marrying states are in the northeast region, but none of the youngest-marrying states are in this region. This could explain the high female breast cancer rate in the northeast region. The data is from Eileen (Eileen, 2014), and link provided in the reference.

	Women	Men
New York	28.8	30.3
Massachusetts	28.8	30.1
Rhode Island	28.5	30.2
Connecticut	28.2	30.0

Table 2: Oldest marrying states

	Women	Men
Utah	23.5	25.6
Idaho	24	25.8

Wyoming	24.5	26.8
Arkansas	24.8	26.3

Table 3: Youngest-marrying states

5. Conclusion

I have analyzed two diseases with geographical preference, these are only small part among all the diseases which have geographical prevalence. Some diseases are terrible not only because the pain it brings to us, but also because it will increase the rate of other diseases, which will add more load both to our body and finance. For example, people with diabetes also likely to have stroke, blindness, kidney disease, heart disease and loss of toes, feet or legs. (Burns, 2018). “Among kidney patients, 75 percent of all new cases re the result of some combination of diabetes and high blood pressure” (Lisa, 2018). Therefore, it is necessary to take some actions to control and decrease the incidence rate.

There is a lot we could do to improve the situation, from precaution to therapy. The most economical way is some education programs aimed at targeting populations, to prevent the occurrence of the disease. For example, we could disseminate relevant information about the disease to increase people’s awareness. Handing out flyers, holding lectures in each community to explain the potential causes and harms of the disease. Advocating people to do more exercise, at the same time, popularizing the public sports facilities. And encouraging people to adjust diet structure, eating more fresh vegetables and fruits, reduce the intake of calories and fat. Once people learn more about this, they will probably pay more attention to it. Early detection of condition is effective to help patients

get treatment in time, this will increase the chance of being cured. However, this requires citizens know much about their health conditions in time. So that governments could hold free examination regularly, this could also help them have more understanding about residents' health status. Once the disease has been detected, some measurements are necessary to prevent further deterioration. Except the patient himself/herself, government also plays an important role in this process. This could help patients especially those don't have access to much resources a lot. Some organizations have realized this and already taken some actions. For example, US Centers for Disease Control and Prevention(CDC) has initiated a five-year program to prevent "heart attacks and strokes by 2017" ("Heart Disease", 2015). And this requires the participant of communities, health systems, nonprofit organizations, federal agencies and private-sector partners ("Heart Disease", 2015).

Above are some general suggestions we could do to change current status, but there is still something we couldn't change, like the race. Some races are more likely to get certain type of diseases than others. For example, African Americans and Hispanics are more likely to have diabetes than non-Hispanic whites (Sandra). Asthma is very common in black children. Their emergency department visit rate, hospitalization rate and death rate are much higher than white children (Russell, 2010). Native Hawaiians and Pacific Islanders have unhealthy lifestyle, they "have higher rates of smoking, alcohol consumption, and obesity" as well as less "access to cancer prevention and control programs" (Russell, 2010). Government should pay more attention to groups which have

this tendency, and the insurance company could also set up some policies for specific type of disease and target population.

In this project, I analyzed the disease prevalence within some regions, and I got some meaningful conclusions. Physical geography of a place affect human life, it not only determines whether humans can live in a certain area or not, it also determines people's lifestyle, as they adapt to the available food and climate patterns. And all of this will have lasting influence on people's health status.

More importantly, I have learned much more from this project. I think this project is more like a summary for my two-years study in SILS. I used what I have learned in courses, researched an area I am interested in, and found something new to me. I would like to specialize in data analytic after graduation, this includes data visualization and data analysis. So I select courses like visual analytics, database, applied statistics and other relevant courses. And my project is also relevant to this, I used what I have learned in the course of applied statistic and visual analytics to visualize this dataset, and analyze from different perspectives. From this project, I also learned that our assumption may out of expectation after experiment, what we could do is to find the reason behind this, and try to explain in other way.

Disease prevalence has relationship with geographical location, and there are many reasons behind this. And there are still a lot to do to change this condition. There are so many people suffering from disease around us. Government could make different

healthcare policies for different states. I think education plays an important role in this. If a child could know the cause and harms for the disease he is likely to get, he must pay attention to it as early as possible.

6. Further work

I have visualized part of the data and done some analysis work, but there are still some limitations and further work to do. Since time and resource is limited, there are still lots of could do with the dataset. I only used twenty attributes in the dataset, but there are more than seventy attributes in total. So that there is still a lot to do with the dataset. It's a useful and meaningful dataset, hope we could find more information behind it to help more people.

References

Adult Obesity in the United States. (n.d.). Retrieved April 04, 2018, from https://stateofobesity.org/adult-obesity/

Advantages of Living in a Big City. (2016, September 07). Retrieved Mar 05, 2018, from https://theclare.com/the-advantages-of-living-in-a-big-city/

Becker, S. (2018, March 24). Cancer in the U.S.: 15 States With the Highest Rates of Diagnoses. Retrieved March 28, 2018, from https://www.cheatsheet.com/culture/cancer-in-the-us-states-with-the-highest-rates-of-diagnoses.html/?a=viewall

Burns, J. (2018, March 07). Cost of Diabetes vs Other Diseases - In the US and Globally. Retrieved March 28, 2018, from https://www.thediabetescouncil.com/cost-diabetes-vs-diseases-us-globally/

Chandra, A. (2005). The Growth Of Physician Medical Malpractice Payments: Evidence From The National Practitioner Data Bank. Health Affairs. doi:10.1377/hlthaff.w5.240

Charles B. Stockdale (2012, March 9). The Six Worst States for Sleep. Retrieved March 28, 2018, from https://247wallst.com/special-report/2012/03/09/the-six-worst-states-for-sleep/2

Diabetes, heart disease, and back pain dominate US health care spending. (2016, December 27). Retrieved March 28, 2018, from https://medicalxpress.com/news/2016-12-diabetes-heart-disease-pain-dominate.html

Diabetes in the United States. (n.d.). Retrieved April 03, 2018, from https://stateofobesity.org/diabetes/

Dugel, P. U., & Tong, K. B. (2011). Development of an Activity-based Costing Model to Evaluate Physician Office Practice Profitability. Ophthalmology, 118(1). doi:10.1016/j.ophtha.2010.04.035

Eileen Shim. (2014, June 27). The Median Age of Marriage in Every State in the U.S., in Two Maps. Retrieved April 05, 2018, from https://mic.com/articles/92361/the-median-age-of-marriage-in-every-state-in-the-u-s-in-two-maps

Hamaty, M. (2015, May 29). How to Manage Your Diabetes in Extreme Summer Heat. Retrieved March 28, 2018, from https://health.clevelandclinic.org/2015/05/how-to-manage-your-diabetes-in-extreme-summer-heat/

Hargens, T., K., E., & B. (2013). Association between sleep disorders, obesity, and exercise: A review. Nature and Science of Sleep, 27. doi:10.2147/nss.s34838

Heart Disease and Stroke Cost America Nearly $1 Billion a Day in Medical Costs, Lost Productivity. (2015, April 29). Retrieved March 28, 2018, from https://www.cdcfoundation.org/pr/2015/heart-disease-and-stroke-cost-america-nearly-1-billion-day-medical-costs-lost-productivity

How much does the U.S. spend to treat different diseases? (2017, May 22). Retrieved March 28, 2018, from https://www.healthsystemtracker.org/chart-collection/much-u-s-spend-treat-different-diseases/#item-circulatory-ill-defined-conditions-check-ups-largest-category-spending

Howe, P. J., & Staff, G. (1995, December 20). Northeast lifestyle tied to breast cancer Rate of fatalities is highest in US. *The Boston Globe (Boston, MA)*. Retrieved April 3, 2018, from http://www.highbeam.com/doc/1P2-8356660.html?refid=easy_hf

Jonathan Blum. (2014, Apr 02). Next Steps in Medicare Data Transparency. *The CMS Blog.* Retrieved March 21, 2018, from https://blog.cms.gov/2014/04/02/next-steps-in-medicare-data-transparency/

Kids' cancer rates highest in Northeast. (2008, June 02). Retrieved March 28, 2018, from http://www.nbcnews.com/id/24920449/ns/health-cancer/t/kids-cancer-rates-highest-northeast/#.WqHxYZPwbVo

Lisa, A. (2018, March 06). 19 Costliest Diseases in the US. Retrieved March 28, 2018, from https://www.gobankingrates.com/saving-money/most-expensive-diseases-us/

Northeast U.S. Has Most Brain Cancer Hospitalization | AHRQ Archive. (2009, March 4). Retrieved March 28, 2018, from https://archive.ahrq.gov/news/newsroom/news-and-numbers/030409.html

Ogara, P. T. (2014). Caution Advised: Medicares Physician-Payment Data Release. New England Journal of Medicine, 371(2), 101-103. doi:10.1056/nejmp1405322

Ornstein, C. (2014, Aug 15). Government Will Withhold One-Third of the Records from Database of Physician Payments. Retrieved Feb 05, 2018, from https://www.propublica.org/article/government-will-withhold-one-third-of-the-records-from-database-of-physicia

Patel, K. P., Masi, D. M., & Brandt, C. B. (2014). Making Sense of the Medicare Physician Payment Data Release: Uses, Limitations, and Potential. doi:10.15868/socialsector.25018

Reschovsky, J. D., Hadley, J., & Romano, P. S. (2013). Geographic Variation in Fee-for-Service Medicare Beneficiaries' Medical Costs Is Largely Explained by Disease Burden. Medical Care Research and Review, 70(5), 542-563. doi:10.1177/1077558713487771

Russell, L. (2010, December 16). Fact Sheet: Health Disparities by Race and Ethnicity. Retrieved March 31, 2018, from https://www.americanprogress.org/issues/healthcare/news/2010/12/16/8762/fact-sheet-health-disparities-by-race-and-ethnicity/

Sandra Gordon.(n.d.), Why Diabetes Is Worse for Latinos and African Americans?. Retrieved April 04, 2018 from https://healthguides.healthgrades.com/stepping-up-your-diabetes-management/why-diabetes-is-worse-for-latinos-and-african-americans

Southern United States. (2018, March 27). Retrieved March 28, 2018, from https://en.wikipedia.org/wiki/Southern_United_States

Steinbrook, R. (2014). Public Disclosure of Medicare Payments to Individual Physicians. Jama, 311(13), 1285. doi:10.1001/jama.2014.1033

Toby Smithson, Alan L. Rubin, Southern Cuisine and Your Diabetic Meal Plan. (n.d.). Retrieved March 28, 2018, from http://www.dummies.com/food-drink/special-diets/diabetes-diets/southern-cuisine-and-your-diabetic-meal-plan/

Whelan, D. (2012, August 05). The 10 Most Expensive Common Medical Conditions. Retrieved March 28, 2018, from https://www.forbes.com/sites/davidwhelan/2012/02/25/the-10-most-expensive-common-medical-conditions/

Where in the Country is Diabetes Belt and Why? | Learn More! (2015, January 13). Retrieved April 03, 2018, from https://www.adwdiabetes.com/articles/where-is-diabetes-belt

Which U.S. States Have the Highest Cancer Rates? (2015, March 26). Retrieved March 28, 2018, from http://blog.dana-farber.org/insight/2015/03/which-u-s-states-have-the-highest-cancer-rates-infographic/

Williams, J. (2017, March 21). Climate Change Effect: Global Warming Could Increase Diabetes. Retrieved March 28, 2018, from http://www.newsweek.com/diabetes-global-warming-climate-change-type-1-type-2-571602

*, M. A., Canavos†, G. C., & Brown, D. M. (2004). An analysis of the variation in billing charges of medical providers: causes and implications. Applied Economics, 36(21), 2377-2384. doi:10.1080/0003684042000286098

Dataset link: https://data.cms.gov/Medicare-Physician-Supplier/Medicare-Physician-and-Other-Supplier-National-Pro/p3uv-6dv4

Data for diabetes rate map: https://stateofobesity.org/diabetes/

Data for cancer rate map: https://statecancerprofiles.cancer.gov/quick-profiles/index.php?statename=wyoming

https://nccd.cdc.gov/uscs/cancersbystateandregion.aspx

Data for cancer table https://www.cdc.gov/cancer/dcpc/data/geographic.htm

Appendix

Screenshots for part of the whole dataset

	A	B	C	D	E	F	G	H	I	J
1	National Pro	Last Name/Organiza	First Name o	Middle Initia	Credentials c	Gender of th	Entity Type c	Street Addre	Street Addre	City of the Pr
2	1245503903	CASSIDY	GEORGE	F	PA-C	M	I	BLDG 1014	27TH ST	FPO
3	1376687954	STALL	LUKE	E	MD	M	I	4700 N LAS	VEGAS BLVD	APO
4	1578743076	NEFF	LUCAS	P	M.D.	M	I	TRAVIS AFB,	DAVID GRAN	APO
5	1629208145	COOMES	MARK	A	M.D.	M	I	23RD MEDIC	3278 MITCHI	APO
6	1659535557	TESSIER	CHARLES		DO	M	I	225 FIRST AVE		APO
7	1750510327	RIAL	NATHANIEL	S	MD/PHD/MF	M	I	NAVAL HOSP	100 BREWST	APO
8	1841458700	LAWSON	TAMARA	D	M.D.	F	I	6900 GEORG	WALTER REE	APO
9	1922088392	STINSON	DARRYL	D	M.D.	M	I	3009 NW	WILSON ST	APO
10	1932363108	ATKINS	JUSTIN	M	M.D.	M	I	CLARK HEALT	BLDG 5-4257	FPO
11	1316177082	FOLARIN	JEAN		NP	F	I	U.S. ARMY H	UNIT 30401	APO
12	1164411658	WEHRER	BRIAN		PA-C, MPAS	M	I	CMR 415, BOX 4449		APO
13	1083924617	VOLNEK	JILLIAN	M	PA-C	F	I	PSC 827	BOX #1000	APO
14	1710214879	FESTEJO	KAREN			F	I	LANDSTUHL	CMR 402	APO
15	1821254236	SCHMIDT	CLAIRE	H	O.D.	F	I	RAF LAKENH	UNIT 5210 B	APO
16	1063855971	PORTER	LAURA	L	ARNP	F	I	2800 BLUE R	REX CARDIO	APO
17	1326161266	SAMPLE	STEPHEN	C	M.D.	M	I	48 MDG/SGOE	UNIT 5210	APO
18	1003015652	DEHQANZADA	ZIA	A	M.D.	M	I	CMR 442	BOX 291	APO
19	1033392659	DUNN	GREGORY	N	MD	M	I	48 MDG	UNIT 5210 B	APO
20	1093867137	SAUNDERS	SHEREE		M.D.	F	I	USS IWO JIM	MEDICAL DE	FPO
21	1144264326	TERRY	MELISSA	V	M.D.	F	I	LRMC	CMR 402	APO
22	1184736795	LIS	THOMAS	S	PA-C, MPH	M	I	PSC 7 BOX 8	APO	APO
23	1194816090	KAPELA	WILLIAM	A	PA-C	M	I	RAF LAKENH	UNIT 5115	APO
24	1295706398	VARONE	RICKY	A	PA-C	M	I	PSC 836	BOX 357	FPO
25	1306038963	BAUMGARTEN	VICTOR	A	PH.D.	M	I	PSC 9 BOX 2884		APO
26	1326139304	LATHAM	JOSHUA	L	D.O.	M	I	48 MDG	UNIT 5210	APO
27	1407814718	BURRIS	WENDELL	G	MD	M	I	USA MEDDA	CMR 411 BLD	APO
28	1487964219	ENSOR	JONELLE	L	PA-C	F	I	CMR 402	LANDSTUHL	APO

A	K	L	M	N	O	P	Q	R	S	T	U
National Pro	Zip Code of t	State Code o	Country Cod	Provider Typ	Medicare Pa	Number of H	Number of S	Number of M	Total Submit	Total Medica	Total Medica
##########	76544	AA	US	Physician Ass	Y	15	132	96	$61,604.00	$9,071.44	$7,052.01
##########	89191	AA	US	Diagnostic Ra	Y	222	2764	1822	##########	$90,714.95	$69,825.63
1578743076	94535	AA	US	General Surg	Y	22	63	49	$46,562.03	$16,487.85	$12,532.77
1629208145	316991500	AA	US	General Surg	Y	31	896	143	$52,482.00	$19,285.66	$9,219.04
1659535557	373892401	AA	US	Family Practi	Y	34	217	71	$16,829.00	$13,313.21	$9,358.99
1750510327	285472538	AA	US	Internal Med	Y	15	133	105	$32,468.00	$16,267.09	$12,753.37
1841458700	203075001	AA	US	Anesthesiolo	Y	48	147	133	##########	$22,788.62	$17,417.01
1922088392	735039042	AA	US	Diagnostic Ra	Y	164	4963	3005	##########	##########	##########
1932363108	28310	AA	US	Emergency M	Y	11	178	118	##########	$19,186.74	$15,005.19
1316177082	91070401	AE	US	Nurse Practit	Y	14	69	37	$6,074.48	$4,686.07	$3,672.50
1164411658	9114	AE	US	Physician Ass	Y	12	22	13	$5,476.50	$1,737.13	$1,364.26
1083924617	96179998	AE	US	Physician Ass	Y	36	1626	201	##########	$53,727.13	$40,847.77
1710214879	9180	AE	US	Physical Ther	Y	6	302	18	$20,845.00	$7,786.99	$6,081.08
1821254236	94610230	AE	US	Optometry	Y	12	40	24	$4,085.00	$3,430.81	$2,381.68
1063855971	9180	AE	US	Nurse Practit	Y	8	102	69	$21,789.00	$6,894.89	$5,283.73
1326161266	94610230	AE	US	Emergency M	Y	39	1042	689	##########	##########	$80,312.75
1003015652	9042	AE	US	General Surg	Y	55	466	256	##########	##########	$82,680.62
1033392659	94610230	AE	US	Anesthesiolo	Y	76	279	214	##########	$30,745.76	$23,538.36
1093867137	95741664	AE	US	Internal Med	Y	30	2033	524	##########	##########	##########
1144264326	9180	AE	US	Obstetrics/G	Y	15	40	24	$13,478.00	$4,333.64	$3,354.77
1184736795	9104	AE	US	Physician Ass	Y	22	577	499	##########	$50,029.76	$34,862.00
1194816090	94615115	AE	US	Physician Ass	Y	76	1490	398	##########	$48,055.48	$32,229.62
1295706398	9636	AE	US	Physician Ass	Y	9	259	218	$64,297.00	$16,622.25	$12,170.96
1306038963	91230029	AE	US	Licensed Clin	Y	6	194	73	$25,728.00	$14,311.34	$10,415.64
1326139304	9461	AE	US	Emergency M	Y	61	842	300	$36,348.60	$19,602.56	$12,594.52
1407814718	9112	AE	US	Family Practi	Y	16	219	129	$35,067.00	$18,893.35	$9,881.41
1487964219	9180	AE	US	Physician Ass	Y	32	1019	309	$65,226.00	$51,670.21	$31,822.28

A	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
National Pro	Total Medica	Drug Suppre	Number of H	Number of D	Number of M	Total Drug Su	Total Drug M	Total Drug M	Total Drug M	Medical Sup	Number of H	Number of M
##########	$8,599.34		0	0	0	$0.00	$0.00	$0.00	$0.00		15	132
##########	$72,272.55		0	0	0	$0.00	$0.00	$0.00	$0.00		222	2764
1578743076	$12,673.96		0	0	0	$0.00	$0.00	$0.00	$0.00		22	63
1629208145	$10,514.43		6	589	54	$10,880.00	$223.35	$167.25	$167.25		25	307
1659535557	$10,350.10		7	68	16	$576.00	$112.24	$95.40	$95.40		27	149
1750510327	$12,860.06		0	0	0	$0.00	$0.00	$0.00	$0.00		15	133
1841458700	$18,265.09		0	0	0	$0.00	$0.00	$0.00	$0.00		48	147
1922088392	##########		0	0	0	$0.00	$0.00	$0.00	$0.00		164	4963
1932363108	$15,447.28		0	0	0	$0.00	$0.00	$0.00	$0.00		11	178
1316177082	$4,349.33	*								#		
1164411658	$1,249.51	*								#		
1083924617	$50,283.46	*								#		
1710214879	$5,607.43		0	0	0	$0.00	$0.00	$0.00	$0.00		6	302
1821254236	$2,581.71		0	0	0	$0.00	$0.00	$0.00	$0.00		12	40
1063855971	$6,523.46		0	0	0	$0.00	$0.00	$0.00	$0.00		8	102
1326161266	$84,561.24		0	0	0	$0.00	$0.00	$0.00	$0.00		39	1042
1003015652	$83,326.71		0	0	0	$0.00	$0.00	$0.00	$0.00		55	466
1033392659	$25,159.38		0	0	0	$0.00	$0.00	$0.00	$0.00		76	279
1093867137	##########		9	364	271	$39,354.00	$37,474.27	$36,370.76	$36,543.91		21	1669
1144264326	$3,573.79		0	0	0	$0.00	$0.00	$0.00	$0.00		15	40
1184736795	$42,190.28		0	0	0	$0.00	$0.00	$0.00	$0.00		22	577
1194816090	$41,567.88		8	360	93	$1,821.80	$857.22	$629.74	$629.74		68	1130
1295706398	$15,541.55		0	0	0	$0.00	$0.00	$0.00	$0.00		9	259
1306038963	$10,637.25		0	0	0	$0.00	$0.00	$0.00	$0.00		6	194
1326139304	$13,964.38		12	389	77	$3,013.60	$355.96	$276.28	$276.28		49	453
1407814718	$10,612.30	*								#		
1487964219	$41,893.52		4	92	31	$2,105.00	$1,756.64	$1,480.74	$1,480.74		28	927

A	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT
National Pro	Number of N	Total Medica	Total Medica	Total Medica	Total Medica	Average Age	Number of B	Number of B	Number of B	Number of B	Number of F	Number of N	Number of N
1245503903	96	$61,604.00	$9,071.44	$7,052.01	$8,599.34	57	54				56	40	46
1376687954	1822	##########	$90,714.95	$69,825.63	$72,272.55	73	317	661	461	383	1085	737	1711
1578743076	49	$46,562.03	$16,487.85	$12,532.77	$12,673.96	70		21	14		26	23	
1629208145	143	$41,602.00	$19,062.31	$9,051.79	$10,347.18	68	34	62	33	14	88	55	
1659535557	71	$16,253.00	$13,200.97	$9,263.59	$10,254.70	69	19	31			42	29	
1750510327	105	$32,468.00	$16,267.09	$12,753.37	$12,860.06	74	13	40	30	22	47	58	72
1841458700	133	##########	$22,788.62	$17,417.01	$18,265.09	67	44	58			65	68	80
1922088392	3005	##########	##########	##########	##########	71	564	1243	871	327	1834	1171	2282
1932363108	118	##########	$19,186.74	$15,005.19	$15,447.28	67	46		31		71	47	78
1316177082						76		17					
1164411658						73							
1083924617						64	79	91			122	79	171
1710214879	18	$20,845.00	$7,786.99	$6,081.08	$5,607.43	85							18
1821254236	24	$4,085.00	$3,430.81	$2,381.68	$2,581.71	68							
1063855971	69	$21,789.00	$6,894.89	$5,283.73	$6,523.46	75		28	23		28	41	56
1326161266	689	##########	##########	$80,312.75	$84,561.24	72	123	242	198	126	415	274	
1003015652	256	##########	##########	$82,680.62	$83,326.71	73	47	84	69	56	135	121	221
1033392659	214	##########	$30,745.76	$23,538.36	$25,159.38	71	46	93	55	20	101	113	198
1093867137	523	##########	##########	##########	$93,214.08	76	39	191	175	119	407	117	
1144264326	24	$13,478.00	$4,333.64	$3,354.77	$3,573.79	55					24	0	
1184736795	499	##########	$50,029.76	$34,862.00	$42,190.28	61	249	132	76	42	262	237	468
1194816090	398	##########	$47,198.26	$31,599.88	$40,938.14	72	56	180	123	39	253	145	359
1295706398	218	$64,297.00	$16,622.25	$12,170.96	$15,541.55	67	63	93	45	17	129	89	
1306038963	73	$25,728.00	$14,311.34	$10,415.64	$10,637.25	63	32	28			45	28	
1326139304	280	$33,335.00	$19,246.60	$12,318.24	$13,688.10	73	30	136	91	43	215	85	276
1407814718						64	52	56			76	53	58
1487964219	309	$63,121.00	$49,913.57	$30,341.54	$40,412.78	75	42	100	106	61	176	133	

Screenshot for part of each state's dataset

	A	B	C	D	E	F	G	H	I
1	Total_charge	Atrial_Fibrillation	Dementia	Asthma	Cancer	Heart_Failur	Kidney_Dise	Obstructive_	Depression
2	71178384.72	24	17	22	13	39	45	30	35
3	22724084.91	11	9	10	9	15	23	12	23
4	19723286.08	6	2	7	12	5	12	7	19
5	15587099.07	9	11	9	8	14	20	12	20
6	12917970.24	17	26	18	10	35	41	22	38
7	11202929.55	19	28	23	12	37	43	31	34
8	10222217	12	10	9	9	19	25	13	16
9	9562620.17	21	35	17	12	36	44	21	36
10	9270386.04	12	8	8	9	17	21	12	13
11	8596446.91	12	10	9	10	15	21	10	23
12	8508509	6	4	9	10	7	15	9	18
13	7309488.23	14	7	10	11	14	19	14	21
14	6899259.87	9	7	12	47	17	27	17	18
15	6748925.5	11	7	12	41	18	31	19	18
16	6739236	11	8	12	47	17	28	18	18
17	6616186.71	9	6	9	7	12	21	13	26
18	6555760.47	22	32	17	12	39	47	23	37
19	6383729.42	7	4	10	7	11	15	14	35
20	6284124.78	21	30	16	13	42	45	29	33
21	5993997.72	18	26	23	12	35	40	31	31
22	5956888	12	6	13	49	16	25	19	17
23	5539528	12	6	13	42	17	26	20	18
24	5518218.96	19	28	17	12	33	42	24	40
25	5497155.1	7	2	9	7	7	11	7	24
26	5485500.54	13	8	12	40	20	34	17	23
27	5270725.15	15	7	7	10	18	29	10	19
28	4736823.93	10	8	13	48	19	32	17	25
29	4530639.5	19	26	22	12	32	42	32	33
30	4507660	10	7	14	42	17	28	17	17